



OPEN

## Cyanophages as an important factor in the early evolution of oxygenic photosynthesis

Ireneusz Ślesak<sup>1</sup> & Halina Ślesak<sup>2</sup>

Cyanophages are viruses that infect cyanobacteria. An interesting feature of many of them is the presence of *psbA* and *psbD*, genes that encode D1 and D2 proteins, respectively. The D1 and D2 are core proteins of the photosystem II (PSII) in cyanobacteria, algae and plants and influence the proper function of oxygenic photosynthesis (OP) in all oxyphototrophs on Earth. The frequent occurrence of *psbA* and *psbD* in cyanophages raises the question whether these genes coevolved with hosts during the early stages of cyanophage and cyanobacterial evolution, or whether they are direct descendants of genes adopted from the genomes of cyanobacterial hosts. The phylogeny of D1/D2 proteins encoded in the genomes of selected cyanophages and oxyphototrophs was reconstructed. In addition, common ancestral sequences of D1 and D2 proteins were predicted for cyanophages and oxyphototrophs. Based on this, the reconstruction of the 3D structures of D1 and D2 proteins was performed. In addition, the ratio of non-synonymous to synonymous ( $d_N/d_S$ ) nucleotide substitutions in the coding sequences (CDSs) of *psbA* and *psbD* was determined. The results of the predicted spatial structures of the D1 and D2 proteins and purifying selection for the CDSs of *psbA* and *psbD* suggest that they belong to the ancient proteins, which may have formed the primordial PSII. It cannot be ruled out that they involved in water oxidation in cyanobacteria-like organisms at early stages of the evolution of life on Earth and coevolved with ancient cyanophages. The data are also discussed in the context of the origin of viruses.

Cyanophages, or bacteriophages (phages) that infect cyanobacteria, were first described in the 1990s<sup>1,2</sup>. To date, the genomes of more than 70 cyanophages, mainly living in marine environments, have been identified and sequenced<sup>3</sup>. Interestingly, the vast majority of described cyanophages mainly infect cyanobacteria from the two genera *Prochlorococcus* and *Synechococcus*, which in turn are the dominant oxyphototrophs on Earth, as they inhabit all oceans<sup>4–7</sup>. Cyanophages are among most abundant ‘semi-life’ forms that play a crucial role in global biogeochemical cycles of carbon, oxygen and nitrogen<sup>8,9</sup>.

A particularly interesting feature of cyanophages is the presence of genes in their genome that encode proteins involved in the photosynthetic activity of cyanobacteria. The best known in this context are the genes coding the PSII core proteins that are most abundant in cyanophages, namely *psbA* for protein D1 and *psbD* for protein D2<sup>3,10,11</sup>. The first photosynthetic gene identified in cyanophages was *psbA*<sup>12,13</sup>. 80% of the cyanophage genomes sequenced so far have *psbA*, and 42% have both *psbA* and *psbD* genes<sup>3</sup>. The function of the cyanophage genes *psbA* and *psbD* is not yet fully understood, but it is likely that they play a key role in infection under changing light conditions that cause photoinhibition of host D1/D2 proteins<sup>14</sup>. Light-induced photoinhibition causes a more rapid degradation of D1/D2 than the completion of the cyanophage developmental cycle. Thus, providing copies of *psbA* (D1) and *psbD* (D2) during viral infection allows the infected cell to be photosynthetically active for as long as is required for full viral particle replication<sup>3,15–17</sup>.

The relatively wide distribution of *psbA* and *psbD* in cyanophages raises the question of whether these genes evolved in co-evolution with hosts in the early stages of cyanophage and cyanobacterial evolution or whether they are direct descendants of genes that cyanophages later adopted from the genomes of cyanobacterial hosts. The answer to this question relates to at least two interrelated issues. First, it concerns the origin of viruses in general and cyanophages in particular. Second, it sheds light on the evolution of oxygenic photosynthesis (OP) too, since cyanobacteria-like organisms are generally considered to be the first oxyphototrophs on Earth<sup>18–20</sup>. According to the most widely held view, the ancestor of cyanobacteria did not possess photosynthetic genes encoding proteins involved in O<sub>2</sub> production. OP probably arose about 2.3–2.6 billion years ago during the

<sup>1</sup>The Franciszek Górski Institute of Plant Physiology, Polish Academy of Sciences, Niezapominajek 21, 30-239 Kraków, Poland. <sup>2</sup>Institute of Botany, Faculty of Biology, Jagiellonian University, Gronostajowa 9, 30-387 Kraków, Poland. email: i.slesak@ifr.pan.edu.pl

so-called Great Oxidation Event (GOE) through lateral gene transfer and the fusion of two photosynthetic systems derived from primordial, non-oxygen evolving bacteria<sup>21–23</sup>. Some recent phylogenetic analyses suggest that OP may have arisen at very early stages of the evolution of life on Earth<sup>24–26</sup>. Therefore, the molecular evolution of proteins at the coding sequence (CDS) level for *psbA* and *psbD* was also studied here in the context of nucleotide substitutions in codons. Most nucleotide substitutions occur at the 3rd position in the codon, less frequently at the 1st and 2nd positions<sup>27</sup>. A nucleotide substitution in a codon can be synonymous and we usually find it at the 3rd position. Generally, synonymous substitutions do not lead to a change in the encoded amino acid in the protein. As a result, the protein retains its biological function. In other words, such a change is neutral and invisible to natural selection. Non-synonymous nucleotide substitutions at the codon lead to a change in the amino acid in the protein. The altered protein can in turn affect the fitness of an organism. Determining the number of non-synonymous ( $d_N$ ) and synonymous ( $d_S$ ) nucleotide substitutions therefore allows us to estimate what type of selection we are dealing with within the CDS under study<sup>27–29</sup>. By comparing the  $d_N/d_S$  ratio, we were thus able to determine which type of selection acts on the CDS for *psbA* and *psbD* in cyanophages and selected oxyphototrophs (see "Methods" for details).

Our data are presented in the context of the various hypotheses that concern origin of viruses. It cannot be ruled out that cyanophages carrying *psbA/psbD* genes are ancient remnants and "living fossils" from the earliest stages of the evolution of life, when a mixture of protocyanobacterial and virus-like replicators existed.

## Results

**Phylogenetic analysis of D1 and D2 proteins.** Phylogenetic trees were reconstructed for D1 and D2 proteins, highlighting cyanophages that have only the gene for the D1 protein (*psbA*) and/or D2 protein (*psbD*) in the genome. In the group of cyanophages examined, no case was found in which only *psbD* was present; when *psbD* was present, it was only together with *psbA* (Fig. 1). In contrast, in some cyanophages *psbA* was present as the only gene for the PSII core protein (Fig. 1).

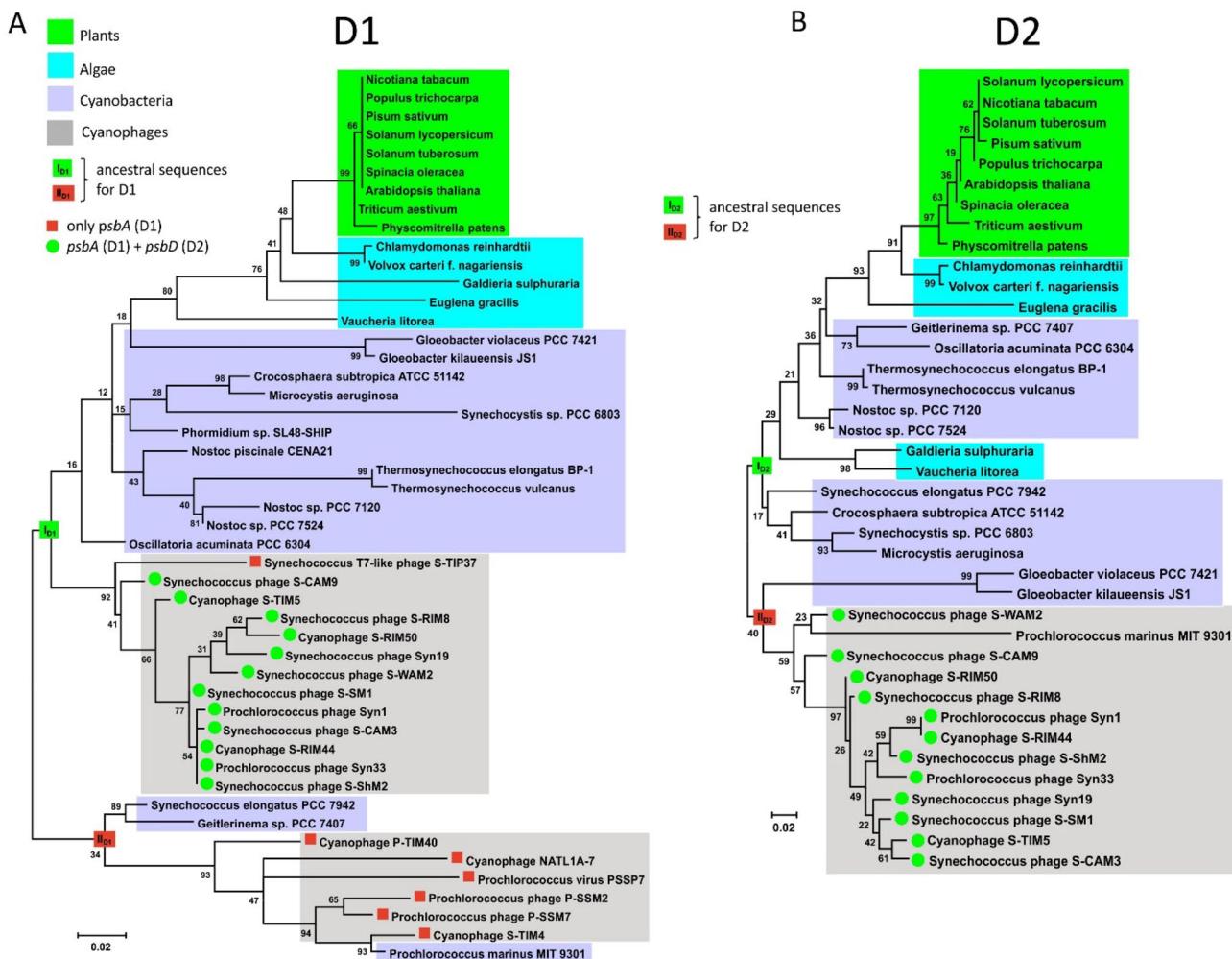
The phylogenetic tree reconstructed for the D1 protein showed several distinct clades: for plants, algae and most of the cyanobacteria analysed (Fig. 1A). One of these clades, with the exception of the *Synechococcus* T7-like phage S-TIP37, includes cyanophages that have *psbD* for the D2 protein in their genome in addition to the *psbA* gene for D1 (Fig. 1A). This entire group of cyanophages most likely shares a common ancestor with D1 for other cyanobacteria, algae and plants.

A separate cyanophage clade is formed by cyanophages that only possess D1. Interestingly, 3 species of cyanobacteria also belong to this clade: *Synechococcus elongatus* PCC 7942, *Geitlerinema* sp. PCC 7407 and *Prochlorococcus marinus* MIT 9301 (Fig. 1). The localisation of the D1 protein in *P. marinus* suggests that *psbA* is derived from cyanophage ancestors of *psbA*. D1 in this group in turn shares a common ancestor with D1 in *S. elongatus* and *Geitlerinema* sp. (Fig. 1A).

**Phylogenetic analysis of D2 protein.** In contrast to the phylogenetic tree for D1, the tree reconstructed for D2 contains fewer protein sequences. This is due to the fact that many fewer cyanophages contain *psbD*, which codes for the D2 protein, in their genome. Moreover, if *psbD* is already present in the cyanophage genome, it is accompanied by *psbA* for the D1 protein (Fig. 1B). Similar to the tree for D1, plants, algae and cyanobacteria form separate clades. However, it is worth noting that D2 occurs in two algal species: *Galdieria sulphuraria* and *Vaucheria litorea* between the D2 sequences for cyanobacteria (Fig. 1B). This result suggests that horizontal gene transfer for *psbD* may have occurred in the past between selected groups of algae and cyanobacteria. This possibility cannot be ruled out, as HGT between bacteria, cyanobacteria, viruses and algae, including the red alga *G. sulphuraria*, has already been observed<sup>30</sup>.

**Ancestral sequence reconstruction and prediction of spatial structure.** The algorithm in the workspace SWISS-MODEL used the D1 template (PDB: 6w1p.1.U), which is 10 amino acids shorter at the N-terminus than the full D1 sequence (UniProt: P0A444, Supplementary Information Notes S5). Homology modelling performed by SWISS-MODEL showed that the predicted and ancestral D1 protein (model\_D1) is 37 and 27 amino acids shorter than the full D1 sequence (UniProt: P0A444) and the D1 template (PDB: 6w1p.1.U) from *Thermosynechococcus elongatus*, respectively. The D1 model lacks the first 26 amino acids and is missing D308 (Fig. 2, Supplementary Information Notes S5). D1 belongs to proteins that play a key role in binding the Mn<sub>4</sub>O<sub>5</sub>Ca cluster<sup>24,25,32</sup>. In *T. elongatus*, the following amino acids of D1 are involved in this binding and the proper function of the oxygen evolving complex (OEC): D170, E189, H332, E333, D342, and A344<sup>33</sup>, <https://www.rcsb.org/3d-view/6W1P>, Supplementary Information Notes S5]. These 6 amino acid residues of the D1 protein are present in all modern organisms that carry out OP<sup>34</sup>. Although the predicted model protein is shorter than the template, namely only shifted by 37 amino acid residues, there is an identical set and pattern of amino acids required for the functioning of OEC: D144, E163, H305, E306, D315 and A317 (Fig. 2, Supplementary Information Notes S5). In general, the identity between the template and the model is ~ 83%. The results indicate that the amino acid sequence of a potentially ancestral D1 protein has not changed significantly compared to the extant D1 proteins found in oxyphototrophs and encoded in cyanophage genomes.

In the case of the D2 protein, the template was also the D2 protein from *T. elongatus* (6dhe.1.D). We obtained a much more similar, almost identical 3D model of the predicted, probably ancestral D2 (Fig. 3, Supplementary Information Notes S6). The differences concern substitutions of only 22 amino acids with similar physicochemical properties (Supplementary Information Notes S6). Although the amino acids of the D2 protein are not directly involved in the binding of the Mn<sub>4</sub>O<sub>5</sub>Ca cluster, they play an important role in proper functioning of PSII<sup>25</sup>. It should be noted that the amino acid residues involved in Mn<sub>4</sub>O<sub>5</sub>Ca binding (Fig. 2) appear to be very conservative and probably occurred in an ancient D1-like protein at a very early stage of OP evolution.

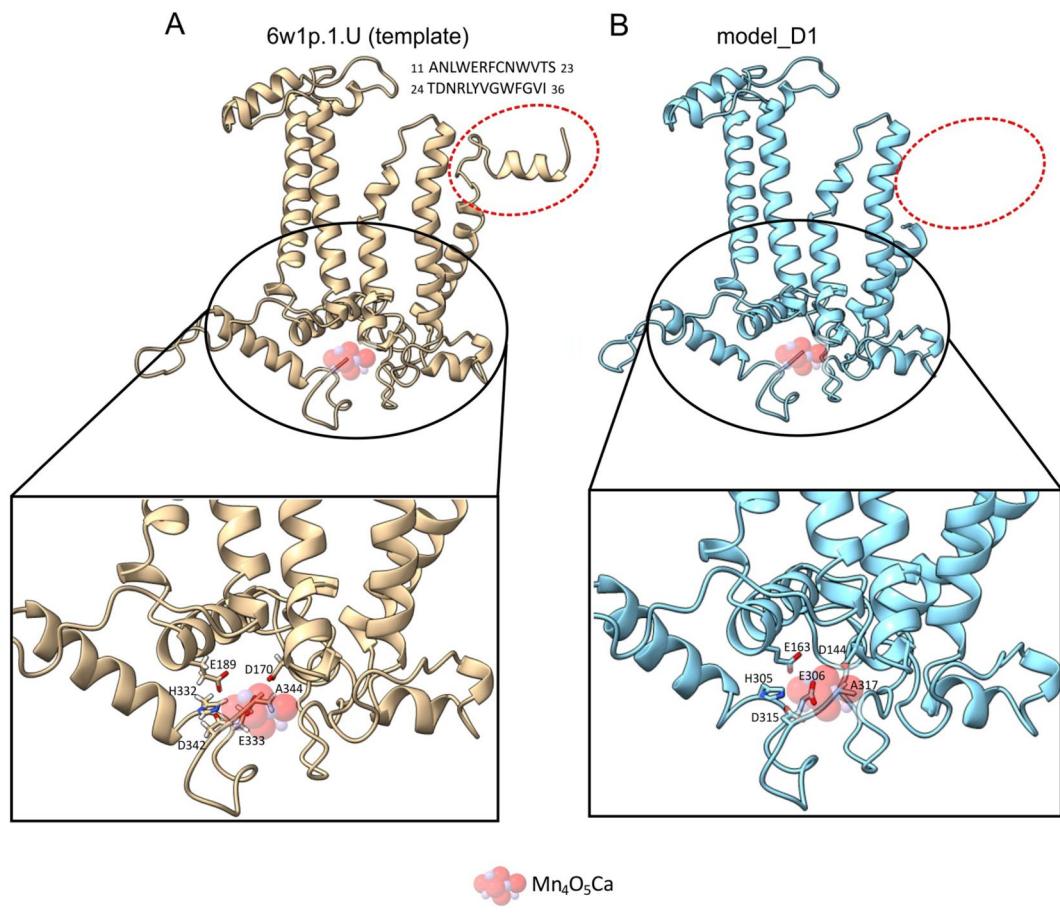


**Figure 1.** Phylogeny of the D1 and D2 proteins for the organism and cyanophage groups analysed in this study (Supplementary Information Notes S1 and Notes S2). **(A)** The unrooted phylogenetic tree of D1. **(B)** The unrooted phylogenetic tree of D2. The deepest nodes for reconstructing the ancestral sequences, designated  $I_{D1}$  and  $II_{D1}$  for D1 **(A)** and  $I_{D2}$  and  $II_{D2}$  for D2 **(B)**, are indicated. The occurrence of *psbA* alone or *psbA* and *psbD* in the cyanophages is indicated. The numbers at the nodes represent bootstrap values (based on 1000 replicates). The scale shows the number of substitutions per amino acid position.

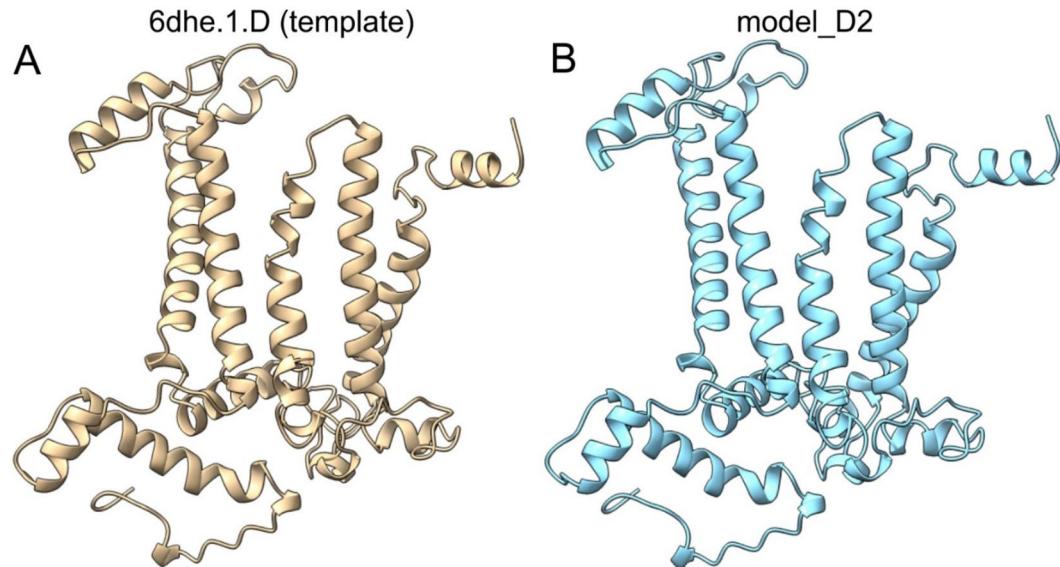
**Determination of  $d_N/d_S$  values.** Analysis of CDSs for *psbA* and *psbD*, encoding D1 and D2 proteins respectively, showed that the  $d_N/d_S$  values for cyanophages, cyanobacteria, algae and plants were lower than 1 ( $d_N/d_S < 1$ ) (Fig. 4, for details see “Methods”). The results show that the CDSs tested are subject to purifying (negative) selection. Interestingly, the  $d_N/d_S$  values for *psbA* and *psbD* were much more similar in both cyanophages and cyanobacteria than the  $d_N/d_S$  values in algae and plants (Fig. 4). These results suggest a close co-evolution of *psbA/psbD* in cyanophages and cyanobacteria.

## Discussion

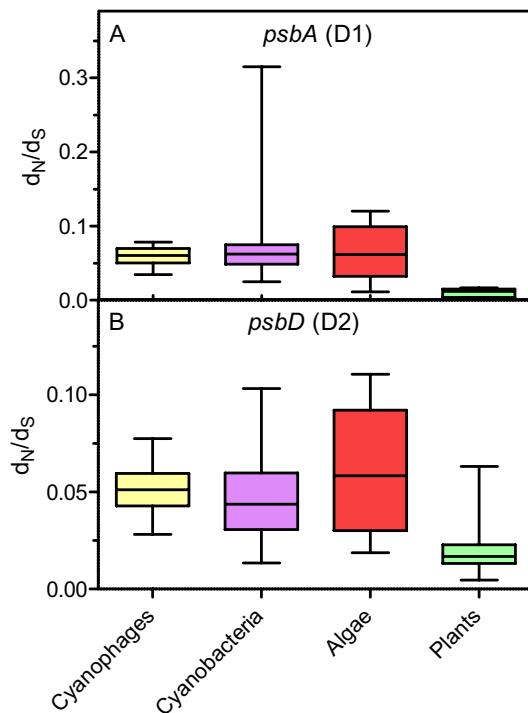
The D1/D2 proteins are homologues formed by duplication of the original gene, and this duplication must have occurred long before the appearance of cyanobacteria<sup>24,25</sup>. The occurrence of early branched cyanobacterial D1 and D2 in clades dominated by cyanophage sequences also suggests an early and continuous gene exchange between cyanophages and their hosts (Figs. 1 and 5). Furthermore, it has already been confirmed that cyanophage genomes encode only D1 or both D1 and D2 proteins. According to our knowledge, no cyanophages have yet been found that have only *psbD* (D2) in their genome<sup>3,10</sup>. It is likely that *psbD* was later lost in some cyanophages. The results also show that purifying selection CDSs of *psbA* and *psbD*, early branching in the trees for D1/D2 proteins, and separate clades with dominance of cyanophage proteins indicate that the *psbA* and *psbD* genes for D1 and D2, respectively, must have evolved very early in both cyanophages and cyanobacteria. It cannot be excluded that the genes *psbA* and *psbD* had a common ancestor prior appearance of cyanobacteria-like organisms. Moreover, the reconstructed spatial structures for the predicted hypothetical ancestral D1 and D2 proteins suggest that they may have been functionally similar enough to the modern forms. The ability to bind the  $Mn_4O_5Ca$  cluster may even have been present in the case of the ancient form of the D1 protein. The common ancestor of the *psbA/psbD* genes and the corresponding D1 and D2 protein products proposed here is



**Figure 2.** The predicted 3D structure of the ancestral D1 protein. (A) The structure of the template sequence from *T. elongatus* (PDB: 6w1p.1.U). (B) The structure of the predicted D1 protein (model\_D1). Shown is the 26 amino acid sequence by which the two structures differ. The Mn<sub>4</sub>O<sub>5</sub>Ca cluster and its binding amino acid residues are shown. The 3D protein structures were visualised and analysed using UCSF ChimeraX software<sup>31</sup>.



**Figure 3.** The predicted 3D structure of the ancestral D2 protein. (A) The structure of the template sequence from *T. elongatus* (PDB 6dhe.1.D). (B) The structure of the predicted D2 protein (model\_D2). The 3D protein structures were visualised and analysed using UCSF ChimeraX software<sup>31</sup>.

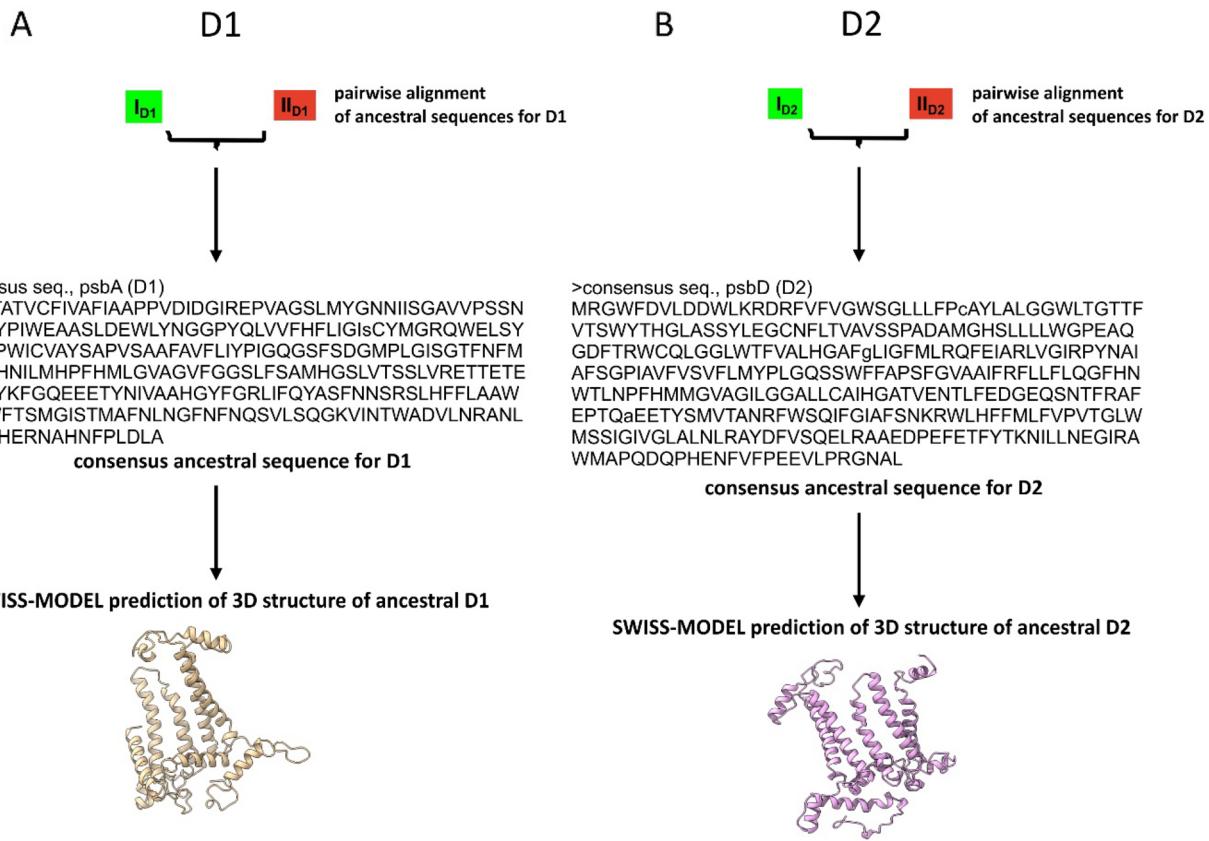


**Figure 4.** The pattern of evolutionary changes in CDSs for  $psbA$  (D1) and  $psbD$  (D2). **(A)**  $d_N/d_S$  values of  $psbA$ . **(B)**  $d_N/d_S$  values of  $psbD$ . Data represent min–max and median values. Values were determined within groups: cyanophages, cyanobacteria, algae and plants. See a detailed list of sequences and species (Supplementary Information Notes S7 and S8).

made plausible by previous data that the D1/D2 heterodimer may have preceded the homodimeric form D0<sup>25</sup>. The predicted structure of the hypothetical D0 protein suggests that it was more similar to the existing D1 than to the D2 sequences, and contained almost the same amino acids as D1, which are responsible for binding the  $Mn_4O_5Ca$  cluster. The only predicted mutation was at position 170, where D replaces E (D170E)<sup>25</sup>. Another important point related to the evolution of OP is also the structural similarity of the Ca-binding site between the homodimeric type I reaction centre (RCI) of green sulphur bacteria and heliobacteria and the Mn-binding site with the D/E residues in PSII<sup>35–37</sup>. Taken together, all these common structural features and the data based on molecular phylogeny provide further evidence for an early evolution of PSII and OEC.

As far as the origin of viruses is concerned, in connection with the phylogeny of  $psbA$  and  $psbD$  in cyanobacteria and cyanophages, one can ask whether it can not only be a rough model for the evolution of cyanobacteria and cyanophages, but whether it can also give an indication of how the evolution of viruses has proceeded in general. In the case of cyanophages and D1/D2 proteins, the results obtained indicate that the primordial hypothesis of viruses in relation to their cellular hosts can be considered<sup>38–41</sup>. Our phylogenetic data showed that both the cyanophage and cyanobacterial D1/D2 proteins are deeply related, suggesting that the cyanophage and cyanobacterial ancestral sequences of  $psbA$  and  $psbD$  arose at least at the same time. This also suggests close co-evolution of  $psbA$  and  $psbD$  between hosts and their cyanophages. Therefore, the data seem to fit, at least in part, the ‘virus early’ hypothesis<sup>38,39,41,42</sup>. However, a certain problem with this virus/cyanophage origin scenario is the question of the origin of the proteins that build their capsids. Viral capsid proteins have no obvious homologues among the proteins of modern cells<sup>43,44</sup>. An analysis of cyanophage capsid proteins is beyond the scope of this study. Nevertheless, following new comparative analyses of capsid proteins, it has been hypothesised that protoviruses, the first replicons or mobile genetic elements, originated with the first cellular forms, i.e. the last universal common ancestor (LUCA), but that the capsid proteins were recruited from later primordial cellular forms<sup>39,44</sup>. All analysed here cyanophages belong to the realm *Duplodnaviria* that possess double-stranded DNA (dsDNA) (data from <https://www.ncbi.nlm.nih.gov>, October 2022). Comparative analyses of viruses and their hosts suggest that *Duplodnaviria*-like dsDNA viruses may have evolved from a population of replicators in the pre-LUCA era, which later gave rise to dsDNA viruses and cellular forms<sup>39</sup>.

At this point, a hypothetical scenario about the origin of viruses and cyanophages in particular is linked to the advent of OP. Cyanophages must have co-evolved with their hosts, as evidenced by the relatively abundant presence of  $psbA$  and  $psbD$  within them. Purifying selection for CDSs of  $psbA/psbD$  has already been described for the major groups of oxyphototrophs<sup>26</sup>. This fact, together with the conserved 3D structure of D1/D2 proteins predicted here, suggests that the ability to bind  $Mn_4O_5Ca$  clusters and to oxidise water may have arisen very early.



**Figure 5.** The scheme of ancestral sequence reconstruction for D1 and D2 proteins. (A) The consensus sequence for D1, obtained from the alignment of the ancestral sequences:  $I_{D1}$  and  $II_{D1}$  from the tree in Fig. 1A (Supplementary Information Notes S3), was used to reconstruct the spatial structure of the protein. (B) The consensus sequence for D2, obtained from the alignment of the ancestral sequences:  $I_{D2}$  and  $II_{D2}$  from the tree in Fig. 1B (Supplementary Information Notes S4), was used to reconstruct the spatial structure of the protein (see “Methods” for details).

## Conclusions

In summary, the D1/D2 proteins and the genomes of the cyanophages that code for them arose in very early stages of the evolution of cyanobacteria and must have been in close relationship with the cyanophages (co-evolution). The *in silico* reconstructed 3D structures of the D1 and D2 proteins and the purifying selection estimated for the CDSs of *psbA* and *psbD* in the studied groups of oxyphototrophs and cyanophages suggest that D1/D2 are ancient proteins that may have formed the primordial PSII that could oxidise water at very early stages of the evolution of life on Earth.

## Methods

**Phylogenetic analysis of D1/D2 proteins.** The molecular phylogenetic analysis was based on D1 and D2 amino acid sequences from selected group of organisms: cyanobacteria, algae, and plants. The latter are defined here as vascular and land plants. Moreover, D1/D2 proteins from cyanophages were also analysed. All sequences were derived from GenBank ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) and UniProt ([uniprot.org](http://uniprot.org)) databases (for details see Supplementary Information Notes S1 and Notes S2). The sequences were aligned with the default parameters of ClustalW implemented in MEGA6<sup>45</sup>. The evolutionary history was inferred by using the Maximum Likelihood method based on the Le and Gascuel<sup>46</sup> model. A discrete Gamma distribution was used to model evolutionary rate differences among sites. The bootstrap consensus tree was inferred from 1000 replicates<sup>47</sup>. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 48 and 39 proteins for D1 and D2, respectively.

**Ancestral sequence reconstruction and prediction of spatial structure.** Based on the phylogenetic trees for D1 and D2 proteins, the deepest nodes were selected, designated  $I_{D1}$  and  $II_{D1}$  for D1 proteins and  $I_{D2}$  and  $II_{D2}$  for D2 proteins (Fig. 1). The ancestral amino acid sequences reconstructed for these nodes were pairwise aligned by the programme Cons (EMBOSS explorer, <https://www.bioinformatics.nl/cgi-bin/emboss/cons>) and the consensus sequences were determined (Fig. 5, Supplementary Information Notes S3 and Notes S4). A 3D structure for D1 and D2 was then predicted for the consensus sequences based on the SWISS-MODELL workspace on March 2022<sup>48</sup>. From the list of known PSII protein structures suggested as templates by the

SWISS-MODEL algorithm, the protein structures of D1 (6w1p.1.U) and D2 (6dhe.1.D) for *T. elongatus* were used. The accession numbers, i.e.: 6w1p and 6dhe correspond to the 3D protein structures of PSII deposited in the Protein Data Bank (PDB) (<https://www.rcsb.org>). The selection was based on the fact that the 3D structures of the D1 and D2 proteins were determined by X-ray diffraction with a resolution of 2.26 and 2.05 Å, respectively<sup>33,49</sup>. The templates and the obtained models were pairwise alignment by EMBOSS needle software (Supplementary Information Notes S5 and Notes S6). The 3D protein structures were visualised and analysed with the software UCSF ChimeraX<sup>31</sup>.

**Coding sequences (CDSs) for *psbA*/*psbD* and determination of the values of non-synonymous ( $d_N$ ) and synonymous ( $d_S$ ) substitutions.** The molecular phylogenetic analysis was based on CDSs for *psbA* and *psbD* taken from selected group of organisms: cyanobacteria, algae, and plants. Moreover, CDSs for these genes from cyanophages were also analysed. All sequences were derived from GenBank ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)), UniProt ([www.uniprot.org](http://www.uniprot.org)), Kyoto Encyclopedia of Genes and Genomes (KEGG, [www.genome.jp/kegg/](http://www.genome.jp/kegg/)), and European Nucleotide Archive (ENA, [www.ebi.ac.uk/ena/browser/home](http://www.ebi.ac.uk/ena/browser/home)) databases for details see Supplementary Information Notes S7 and Notes S8). The number of non-synonymous ( $d_N$ ) and synonymous ( $d_S$ ) nucleotide substitutions was determined in the analysed CDSs. The ratio of  $d_N$  to  $d_S$  ( $d_N/d_S$ ) was calculated to estimate the type of selection. According to the  $d_N/d_S$  method,  $d_N/d_S = 1$  means neutral evolution,  $d_N/d_S < 1$  indicates purifying (negative) selection, and  $d_N/d_S > 1$  means positive selection<sup>29</sup>. The sequences were aligned using the default parameters of ClustalW implemented in MEGA6<sup>45</sup>. The  $d_N$  and  $d_S$  values were calculated using the software DnaSP v. 6.12.03<sup>50</sup>, then  $d_N/d_S$  values were determined.

## Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 19 August 2022; Accepted: 21 November 2022

Published online: 29 November 2022

## References

1. Suttle, C. A. & Chan, A. M. Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: Abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92**, 99–109 (1993).
2. Wilson, W. H., Joint, I. R., Carr, N. G. & Mann, N. H. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH7803. *Appl. Environ. Microbiol.* **59**, 3736–3743 (1993).
3. Puxty, R. J., Millard, A. D., Evans, D. J. & Scanlan, D. J. Shedding new light on viral photosynthesis. *Photosynth. Res.* **126**, 71–97 (2015).
4. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**, 83–86 (2007).
5. Rohwer, F. & Thurber, R. V. Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009).
6. Flombaum, P. *et al.* Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci. USA* **110**, 9824–9829 (2013).
7. Fridman, S. *et al.* A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected *Prochlorococcus* cells. *Nat. Microbiol.* **2**, 1350–1350 (2017).
8. Suttle, C. A. Marine viruses—major player in the global ecosystems. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
9. Mann, N. H. & Clokie, M. R. J. Cyanophages. In *Ecology of Cyanobacteria II* (ed. Whitton, B. A.) 535–557 (Springer, 2012).
10. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4**, e234 (2006).
11. Mazor, Y. *et al.* The evolution of photosystem I in light of phage-encoded reaction centres. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 3400–3405 (2012).
12. Lindell, D. *et al.* Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. USA* **101**, 11013–11018 (2004).
13. Millard, A., Clokie, M. R. J., Shub, D. A. & Mann, N. H. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc. Natl. Sci. USA* **101**, 11007–11012 (2004).
14. Bailey, S., Clokie, M. R. J., Millard, A. & Mann, N. H. Cyanophage infection and photoinhibition in marine cyanobacteria. *Res. Microbiol.* **155**, 720–725 (2004).
15. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–89 (2005).
16. Mella-Flores, D. *et al.* *Prochlorococcus* and *Synechococcus* have evolved different adaptive mechanisms to cope with light and UV stress. *Front. Microbiol.* **3**(285), 1–20 (2012).
17. Jacobson, T. B., Callaghan, M. M. & Amador-Noguez, D. Hostile takeover: How viruses reprogram prokaryotic metabolism. *Annu. Rev. Microbiol.* **75**, 515–539 (2021).
18. Schirrmeister, B. E., Sánchez-Baracaldo, P. & Wacey, D. Cyanobacterial evolution during the Precambrian. *Int. J. Astrobiol.* **15**, 187–204 (2016).
19. Sanchez-Baracaldo, P. & Cardona, T. On the origin of oxygenic photosynthesis and cyanobacteria. *New Phytol.* **225**(4), 1440–1446 (2019).
20. Ślesak, I., Kula, M., Ślesak, H., Miszalski, Z. & Strzałka, K. How to define obligatory anaerobiosis? An evolutionary view on the antioxidant response system and the early stages of the evolution of life on Earth. *Free Radic. Biol. Med.* **140**, 61–73 (2019).
21. Hohmann-Marriott, M. F. & Blankenship, R. E. Evolution of photosynthesis. *Ann. Rev. Plant Biol.* **62**, 515–548 (2011).
22. Shih, P. M. *et al.* Crown group oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19–29 (2017).
23. Soo, R. M. *et al.* On the origins of oxygenic photosynthesis and aerobic respiration in cyanobacteria. *Science* **355**, 1436–1440 (2017).
24. Cardona, T., Sánchez-Baracaldo, P., Rutherford, A. W. & Larkum, A. W. Early Archean origin of photosystem II. *Geobiology* **17**, 127–150 (2019).
25. Oliver, T., Sánchez-Baracaldo, P., Larkum, A. W., Rutherford, A. W. & Cardona, T. Time-resolved comparative molecular evolution of oxygenic photosynthesis. *Biochim. Biophys. Acta Bioenerg.* **1862**, 6 (2021).
26. Ślesak, I., Mazur, Z. & Ślesak, H. Genes encoding the photosystem II proteins are under purifying selection: An insight into early evolution of oxygenic photosynthesis. *Photosynth. Res.* **153**, 163–175 (2022).

27. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford University Press, 2000).
28. Saitou, N. Natural selection. In *Introduction to Evolutionary Genomics. Computational Biology* (eds Dress, A. et al.) 125–154 (Springer, 2013).
29. Pybus, O. G. & Shapiro, B. Natural selection and adaptation of molecular sequences. In *The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (eds Lemey, P. et al.) 407–418 (Cambridge University Press, 2009).
30. Lee, J. et al. Reconstructing the complex evolutionary history of mobile plasmids in red algal genomes. *Sci. Rep.* **6**, 23744 (2016).
31. Pettersen, E. F. et al. UCSF Chimera X: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
32. Nixon, P. J., Sarcina, M. & Diner, B. A. The D1 and D2 core proteins. In *Photosystem II: The Light-Driven Water: Plastoquinone Oxidoreductase* (eds Wydrzynski, T. & Satoh, K.) 71–93 (Springer, 2005).
33. Ibrahim, M. et al. Untangling the sequence of events during the S2 → S3 transition in photosystem II and implications for the water oxidation mechanism. *Proc. Natl. Acad. Sci. USA* **117**, 12624–12635 (2020).
34. Shen, J. The structure of photosystem II and the mechanism of water oxidation in photosynthesis. *Ann. Rev. Plant Biol.* **66**, 23–48 (2015).
35. Cardona, T. & Rutherford, A. W. Evolution of photochemical reaction centres: More twists? *Trends Plant Sci.* **24**, 1008 (2019).
36. Gisriel, C. J., Azai, C. & Cardona, T. Recent advances in the structural diversity of reaction centers. *Photosynth. Res.* **149**, 329–343 (2021).
37. Shimada, Y. et al. Post-translational amino acid conversion in photosystem II as a possible origin of photosynthetic oxygen evolution. *Nat. Commun.* **13**, 4211 (2022).
38. Koonin, E. V. *The Logic of Chance. The Nature and Origin of Biological Evolution* (Pearson Education, 2012).
39. Krupovic, M., Dolja, V. V. & Koonin, E. V. Origin of viruses: Primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).
40. Nasir, A., Romero-Severson, E. & Claverie, J.-M. Investigating the concept and origin of viruses. *Trends Microbiol.* **28**, 959–967 (2020).
41. Harris, H. M. B. & Hill, C. A place for viruses on the tree of life. *Front. Microbiol.* **11**, 3449 (2021).
42. Krupovic, M., Dolja, V. V. & Koonin, E. V. The LUCA and its complex virome. *Nat. Rev. Microbiol.* **18**, 661–670 (2020).
43. Cheng, S. & Brooks, C. L. 3rd. Viral capsid proteins are segregated in structural fold space. *PLOS Comput. Biol.* **9**, e1002905 (2013).
44. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. USA* **114**, E2401–E2410 (2017).
45. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
46. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
47. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution (NY)* **39**, 783–791 (1985).
48. Waterhouse, A. et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**(W1), W296–W303 (2018).
49. Kern, J. et al. Structures of the intermediates of Kok's photosynthetic water oxidation clock. *Nature* **563**, 421–425 (2018).
50. Rozas, J. et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).

## Acknowledgements

This research was supported by The Franciszek Górski Institute of Plant Physiology Polish Academy of Sciences in Kraków, Poland.

## Author contributions

Conceptualization, methodology, data collection and analysis, writing-preparation of the original draft: I.Ś. Data analysis, writing-review and editing: H.Ś.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-24795-1>.

**Correspondence** and requests for materials should be addressed to I.Ś.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022